한국산업정보학회논문지 제10권 제1호(2005. 3)

퍼지추론 기반 대표 키워드 추출방법의 성능 평가†

(Performance Evaluation of the Extraction Method of Representative Keywords by Fuzzy Inference)

노순억*, 김병만*, 오상엽*, 이현아*
(Sun-Ok Rho, Byeong Man Kim, Sang Yeop Oh, Hyun Ah Lee)

요 약 본 논문은 퍼지 추론을 이용하여 소수문서로부터 대표 용어들을 추출하고 가중치를 부여하는 기존 방법의 유용성을 평가하고자 GIS (Generalized Instance Set) 알고리즘에 이를 적용시켜 그 성능을 평가하여 보았다. GIS 는 학습 문서 집합에 대한 일반화 (generalization) 과정을 통해 문서 그룹들을 형성하고 이 그룹의 대표 문서 (generalized instance)를 생성한 후 k-NN 알고리즘을 적용하는 방법이다. 본 논문에서는 바로 이 일반화 과정의 한 방법으로 퍼지 추론을 이용한 방법을 사용하였다. 상대적 성능 평가를 위하여 이 일반화(generalization) 과정에 Rocchio와 Widrow-Hoff 방법도 적용시켜 문서 분류 성능을 비교하였다. 실험 결과, 긍정적 문서만을 고려할 경우는 좋은 성능을 보이지만 부정적 문서를 같이 고려할 경우는 성능이 상대적으로 좋지 않음을 확인 할 수 있었다.

핵심주제어 : 사용자 모델링, 관심사 추출, 문서분류, 키워드 추출

Abstract In our previous works, we suggested a method that extracts representative keywords from a few positive documents and assigns weights to them. To show the usefulness of the method, in this paper, we evaluate the performance of a famous classification algorithm called GIS (Generalized Instance Set) when it is combined with our method. In GIS algorithm, generalized instances are built from learning documents by a generalization function and then the K-NN algorithm is applied to them. Here, our method is used as a generalization function. For comparative works, Rocchio and Widrow-Hoff algorithms are also used as a generalization function. Experimental results show that our method is better than the others for the case that only positive documents are considered, but not when negative documents are considered together.

Key Words : User Modeling, Extraction of User Preference, Document Classification, Keywords Extraction

1. 서 론

웹 검색 엔진 혹은 다양한 정보 검색 시스템을 이용하는 일반 사용자는 자신이 원하는 내용에 가장 적합한 정보를 찾고자 관심 대상 영역에 대한 제한된 어휘력과 전문성을 바탕으로 검색 질의어를 구성한다. 마찬가지로 정보 필터링 시스템 이용 시 사용자는 적절한 정보를 추천 혹은 제공받고자 자신의 프로파일에 관심 사항을 기술한다.

검색 시스템의 경우 제공된 질의어로 검색 기능을 수행하고 그 결과에 대해서 사용자로부터 피드백을 받거나 검색된 결과를 이용해 자동으로 질의어를 수정하고 중요도를 재산정하는 등의 부가 기능들을 수

† 본 연구는 금오공과대학교 학술연구비에 의하여 연구되었음.
* 금오공과대학교 컴퓨터공학부

행함으로써 사용자에게 편의성을 제공하고 검색 효율을 높이고 있다. 정보 필터링 시스템 역시 위와 비슷한 성격의 프로파일 수정 과정들을 가진다.

사용자에 의한 프로파일 관리는 사용자에게 부담을 줄 수 있고 용어 불일치 문제로 인한 부적절한 필터링 결과를 가져 올 수 있다. 따라서 사용자의 프로파일 작성의 부담과 용어 불일치 문제의 수위를 낮추기 위해서 사용자로부터 관심 내용과 유사한 문서 집합을 제공받아 이를 활용하는 것도 하나의 해결 방법이 될 수 있다. 관심 문서를 제공받는 방법은 사용자로부터 명시적으로 입력을 받을 수도 있고 간접적으로 사용자가 관심을 갖는 문서를 파악하여 이를 사용할 수도 있다. 이 경우에 발생되는 문제점은 제공된 문서 집합으로부터 사용자를 대신해서 대표 용어를 추출하고 이들에게 어느 정도의 중요도를 부여할 것인가이다. 이러한 접근 방법의 하나로 본 연구자는 퍼지 추론을 이용한 방법을 제시하였다 [2].

[2]에서는 학습 문서에 퍼지 추론을 적용시켜 초기 중요 단어들을 추출하고 이 단어들을 초기 질의어로 간주하여 기존의 질의 자동 확장 방법을 적용시켜 중요도를 재산정하는 방법을 제안하였고 실험을 통해 그 유용성을 확인하였다. 하지만 [2]의 방법은 사용자가 피드백하는 문서가 소수이며 긍정적인 문서만 피드백한다는 가정 하에서 제안되었다. 따라서, 본 논문에서는 [2]의 방법을 다수의 문서가 주어진 경우와 부정적 문서도 같이 주어진 경우로 확장하여 그 유용성을 확인하고자 하였다. 이를 위해 GIS (Generalized Instance Set) 알고리즘 [10, 11]에 [2]의 방법을 적용시켜 그 성능을 평가하였다.

GIS 는 학습 문서 집합에 대한 일반화(generalization) 과정을 통해 문서 그룹들을 형성하고 이 그룹의 대표 문서 (generalized instance)를 생성한 후 k-NN 알고리즘 [5]을 적용하는 방법이다. 본 논문에서는 바로 이 일반화 과정의 한 방법으로 [2]의 방법을 사용하였다. 그리고, 이 일반화(generalization) 과정에 Rocchio와 Widrow-Hoff 방법도 적용시켜 상대적 성능 비교를 하였다.

2. 관련 연구

본 논문에서 다루는 내용은 문서 자동 분류에서 문서 범주의 대표용어를 구성하는 문제와 유사하다. 따라서, 이번 장에서는 다양한 문서 분류 방법 - Decision tree, Decision rule, Neural network, Rocchio, Widrow-Hoff, k-NN, GIS, SVM 등 [1, 3, 4, 5, 6, 7, 9] - 중에서 학습 문서 집합의 중심 벡터를 구성하는 분류 방법인 Rocchio와 Widrow-Hoff 그리고 이들을 이용하는 GIS(Generalized Instance Set) 방법[3, 10, 11]에 대해서 살펴본다.

2.1 선형 분류기 (Linear classifier)

정보 검색 시스템(IR system)들은 일반적으로 문서를 $X=(x_1, x_2, \ldots, x_d)$와 같이 특징들의 벡터로 표현한다. 여기서 x_j는 해당 문서에서 특징 j 가 가지는 가중치 값을 의미하고 d 는 특징들의 개수를 나타낸다. 전체 문서 집합으로부터 불용어 처리 및 스테밍 과정을 거쳐 추출된 용어들을 특징들로 삼을 수 있고 가중치는 해당 문서 내에서 특징의 발생 빈도수나 역문헌 빈도수등을 이용하여 계산할 수 있다.

문서 검색 시스템의 경우 보통 d 개의 입력 변수들을 가진 함수 f 를 각각의 벡터 X 에 적용하여 그 결과값 $f(X)$를 기반으로 검색된 문서들을 순서화 한다. 문서 분류 시스템은 유사한 방식으로 $f(X)$ 값을 계산하여 특정 범주의 소속 정도를 부여한다. 이러한 함수들은 선형적(linear)이며 아래와 같은 가중치 벡터 W 와 특징 벡터 X 의 내적(dot product)으로 표현 될 수 있다 [1].

$$f(X) = W \cdot X = \sum_{j=1}^{d} w_j x_j$$

문서 분류기들 중 Rocchio, Widrow-Hoff 선형 분류기들은 학습 문서 집합을 대상으로 학습 과정을 거쳐 새로운 문서를 정확하게 분류할 수 있는 가중치 벡터 W 를 유도한 뒤 이를 사용한다.

2.1.1 Rocchio 분류기

Rocchio 분류기 [1, 3]는 벡터 공간 모델에서 연관성 피드백을 위한 Rocchio 식을 문서 분류에 적용한 것으로 아래 식 1을 사용하여 배치모드(batch

mode)로 문서 집합의 중심 벡터를 계산한다.

$$w_j = \alpha\, w_{1,j} + \beta \frac{\sum_{i \in C} x_{i,j}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{i,j}}{n - n_C} \quad \text{(식 1)}$$

여기서, $x_{i,j}$는 문서 i에서의 단어 j의 가중치를, n은 학습 문서(training documents)의 수를, C는 긍정적 문서(positive documents)집합을, n_C는 긍정적 문서의 수를 나타낸다. 그리고, 조정값 α, β 그리고 γ 는 각각 초기 가중치 벡터, 긍정적 문서 벡터들 그리고 부정적 문서 벡터들의 상대적 영향력을 제어하기 위해 사용된다. α=0, β=1, γ=1 일 경우 긍정적 문서들에 대한 가중치의 평균과 부정적 문서들에 대한 가중치 평균의 차이가 문서 집합에 대한 해당 특징의 가중치 값으로 계산된다. 위 식을 통해 생성된 분류기는 양수의 가중치 값을 가지도록 제한되어 아래와 같이 값 w'_j을 사용한다.

$$w'_j = \begin{cases} w_j, \text{ if } w_j > 0 \\ 0 \quad otherwise \end{cases}$$

2.1.2 Widrow-Hoff 분류기

아래 식 2는 Widrow-Hoff 분류 방법 [1]으로서 한 단계마다 하나씩 문서를 처리하면서 중심 벡터의 가중치들을 갱신하고 있으며 온라인(on-line) 환경에서 사용될 수 있다. 초기 벡터는 보통 $W_1 = (0, \ldots, 0)$ 으로 설정되며 각 단계에서의 새로운 가중치 벡터 W_{i+1} 는 라벨 y_i를 가진 학습 문서 벡터 X_i 를 사용하여 이전 가중치 벡터 W_i로부터 계산되어진다. 라벨 y_i 는 해당 문서가 긍정적 문서일 경우 1, 부정적 문서일 경우 0 의 값을 각각 가진다. 식 2를 살펴보면 단계마다 기존 가중치 벡터를 이용한 분류 결과에 대하여 오차를 줄여 나가는 방향으로 가중치 값을 조정함을 볼 수 있다.

$$w_{i+1,j} = w_{i,j} - 2\eta(W_i \cdot X_i - y_i) x_{i,j} \quad \text{(식 2)}$$

여기서, η는 학습율(learning rate)을 나타낸다.

2.2 k-NN (k nearest neighbors) 분류기

k-NN 분류기는 문서 분류 문제를 위한 Expert Network(ExpNet) [5] 이란 문서 분류 시스템에서 사용된 바 있으며 좋은 성능을 보여주었다. k-NN는 선형 분류기들처럼 학습 과정을 통해 해당 범주를 대표하는 가중치 벡터와 같은 분류 체계를 미리 유도해 내지는 않는다. 대신에 학습 문서들을 분류 대상 문서와의 코사인 유사도 값들에 따라 내림차순으로 정렬한 뒤 상위 k 개의 문서들을 선택한다. 이렇게 선택된 k 개의 문서들과 해당 범주와 관련성을 고려하여 분류 대상 문서 X의 소속 정도를 계산한다. 즉, 아래와 같은 식을 통하여 X가 범주 C에 속할 정도가 계산되어 진다. 이 값이 임계값보다 크면 X가 범주 C에 소속되는 것으로 판단한다.

$$\text{rel}(C, X) = \sum_{D_j \in D} \text{sim}(X, D_j) \times Assoc(C, D_j)$$

여기서, D는 학습 문서중 문서 X와 유사도가 높은 k개의 문서집합을 의미하며 $Assoc(C, D_j)$는 문서 D_j와 범주 C와의 관련 정도로 ExpNet에서는 아래와 같이 산정하여 사용한다.

$$Assoc(C, D_j) = \frac{n_1}{n_2}$$

여기서, n_1은 학습 문서 집합에서 문서 D_j가 범주 C에 할당된 횟수를, n_2는 문서 D_j가 학습 문서 집합 내에서 나오는 회수를 의미한다. 학습 문서 집합 내에서 문서 D_j가 한번만 나온다고 가정하면 그 문서가 C로 분류되어 있는 경우 1이고 그렇지 않은 경우 0이 된다.

2.3 GIS(Generalized Instance Set) 분류기

GIS(Generalized Instance Set) 분류기 [10, 11]는 선형 분류기와 k-NN (k nearest neighbors) 분류기의 장.단점을 고려하여 두 방법을 결합한 분류기이다. GIS 분류기에서는 GIS 생성 알고리즘을 사용하여 적절한 긍정적 문서들과 부정적 문서들을 선택하여 노이즈를 제거한 후 Rocchio, Widrow-Hoff 등과 같은 선형 분류기를 적용시켜 대표문서

(generalized instance) 집합을 구축한다.

대표문서들이 생성되고 나면 이를 학습 문서처럼 취급하는데 아래와 같은 k-NN 방법과 유사한 방법을 적용하여 문서를 분류한다.

$$Score(X, C) = \sum_{G \in GS} \mathrm{Sim}(G, X) \times Assoc(G, C)$$

$$Assoc(G, C) = \frac{P_k}{P}$$

여기서, $Assoc(G, C)$ 는 generalized instance G 와 범주 C와의 연관성을 나타내는 함수로 P 는 학습 문서 집합에서 범주 C에 속하는 긍정적 문서 개수를 의미하고, P_k 는 범주 C에 속하는 긍정적 문서 중에서 G의 상위 k 인접 문서 집합에 속하는 문서의 개수를 의미한다. 즉, $Assoc(G, C)$ 는 C의 긍정적 문서 집합이 G를 구성하는데 얼마나 관여했는가를 나타내는 척도로 해석할 수 있다. 예를 들어, 대표문서 G를 구성할 때 관여한 문서가 모두 범주 C에 속한다면 $Assoc(G, C)$는 1이고 반대로 범주 C에 속한 문서가 하나도 G를 구성하는데 관여하지 않았다면 0이다.

대표문서 집합내의 각각의 G 와 범주 C 에 대한 $Assoc(G, C)$ 이 구해지면 $Score(X, C)$ 계산식을 통해 최종적으로 범주 C 에 대한 새로운 문서 X 의 소속 정도가 구해지게 된다. 이렇게 계산된 소속 정도 값은 미리 정의된 임계값(threshold)과의 비교를 통해서 소속 여부를 결정하는데 사용된다. 한마디로 GIS 분류기는 일반문서 대신에 대표문서에 k-NN 방법을 적용시키는 것으로 해석하면 된다.

3. 퍼지추론을 이용한 소수 관련문서의 대표 단어 추출

퍼지 추론을 이용한 소수 문서의 대표 키워드 추출 (이하 RKEF) [2]에서는 사용자가 제시한 소수의 예제 문서집합으로부터 사용자의 관심사항을 가장 잘 대변하는 대표 용어들을 추출하고 이들의 가중치를 부여하는 문제를 다루고 있다. 이는 앞에서 언급한 Rocchio와 Widrow-Hoff 방법들이 학습 문서 집합을 대표하는 중심 벡터를 구성하는 것과 성격이 유사하므로 문서분류 문제에 이 방법을 적용할 수 있다(식 4 참조).

RKEF에서 제시한 소수 예제 문서 집합의 대표 벡터를 구성하는 과정은 크게 3 단계로 다음과 같이 요약할 수 있다.

- 퍼지 추론을 이용한 대표 용어 중요도 계산
- 초기 대표 용어 선택
- 용어 가중치 재산정과 대표용어 자동확장

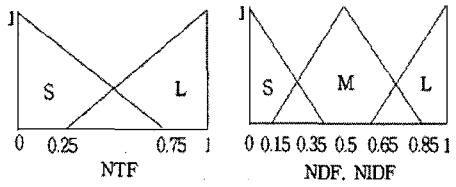
3.1 퍼지 추론을 이용한 대표 용어 중요도 계산

아래 식 3은 퍼지 추론에 사용된 퍼지 입력 변수들의 입력값을 계산하기 위한 식들을 나타내고, 그림 1, 2, 3은 퍼지 입.출력 변수들과 추론에 사용된 규칙들을 각각 나타내고 있다 (퍼지 추론 부분에 관한 자세한 설명은 [2] 참조).

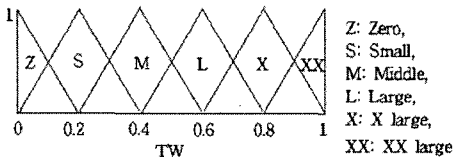
$$NTF_i = \frac{TF_i / DF_i}{\max_j [TF_j / DF_j]},$$

$$NDF_i = \frac{DF_i / TD}{\max_j [DF_j / TD]}, \qquad \text{(식 3)}$$

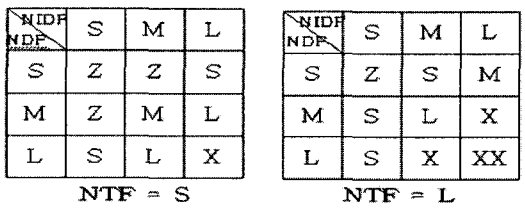
$$\mathrm{NIDF}_i = \frac{IDF_i}{\max_j [IDF_j]}$$



[그림 1] 퍼지 입력변수들



[그림 2] 퍼지 출력변수들



[그림 3] 퍼지 추론 규칙

3.2 초기 대표 용어 선택

퍼지 추론을 통해 후보 용어들의 중요도 값들이 계산되어지면 이 값들에 따라 선택 우선 순위를 부여한다. 각각의 관심 문서가 적어도 1개의 초기 대표 용어를 포함해야 한다는 제약 사항이 있다. 그림 4는 초기 대표 용어 선택 알고리즘을 보여주고 있다. 초기 대표 용어의 개수는 후보 용어와의 발생 빈도 유사도 계산의 정확성을 위해 최소로 유지하면서 위의 제약 사항에 따라 용어들을 선택한다.

3.3 용어 가중치 재산정과 대표용어 자동확장

식 4는 RKEF 에서 사용된 계산식으로서 사용자가 제시한 소수 예제 문서집합에 대해서 이를 대표하는 벡터를 구성하는 최종 계산식을 보여주고 있다 [2]. 식 5는 퍼지 추론 과정에 의해 선택된 핵심적인 초기 대표 용어들과 다른 일반 용어들간의 발생 빈도수의 유사성을 계산하기 위해 사용된다. 핵심 용어들과의 발생 빈도수 유사 정도에 따라 용어의 가중치 값을 조정하게 하는 역할을 한다.

$$w_j = \sum_{i \in C} (x_{i,j} \times RD_{i,j}(K)) \qquad \text{(식 4)}$$

$$RD_{i,j}(K) = 1 - \log_p \left(\sqrt{\frac{\sum_{k=1}^{n} (TF_{i,k} - TF_{i,j})^2}{n}} \right) \qquad \text{(식 5)}$$

여기서, w_j는 대표 벡터에서 용어 j 의 가중치를, $x_{i,j}$는 문서 i 에서 용어 j 의 가중치를, K는 퍼지 추론으로 선택된 초기 대표 용어 집합을, $RD_{i,j}(K)$는 문서 i 에서 용어 j 와 용어 집합 K 간의 관련정도를, $TF_{i,k}$는 문서 i 에서 초기 대표 용어 k 의 발생 빈도수를, $TF_{i,j}$는 문서 i 에서 용어 j 의 발생 빈도수를, n은 초기 대표 용어 수를 나타낸다.

4. 다수의 관련문서와 비관련문서를 고려할 경우의 적용 방법 및 성능 평가

RKEF의 유용성을 평가하기 위해 다양한 실험을 하였다. RKEF 방법이 원래 소수의 관심문서가 주어진 경우에 맞추어 제안된 방법이기 때문에 먼저, 소수의 긍정적 문서가 주어진 경우의 Roccio와 Widrow-Hoff 방법과의 성능을 비교하였다. 그리고, 다수의 긍정적 문서가 주어졌을 경우의 적용 방안과 그 성능 평가를 하였다. 마지막으로 부정적 문서까지 고려할 경우의 적용 방안과 성능 평가도 하였다.

```
Input: DS ( Example Documents Set )
       TS ( Candidate Terms Set )

1] Procedure get_ITS(DS, TS)
2] ITS: Initial Representative Terms Set,
       initialized to empty.
3] TS': Temporary Terms Set, initialized to TS.
4] d, t: Document and Term element respectively.
5] Repeat
6]   Select a document element as d from DS.
7]   Repeat
8]     Select the highest element as t in TS'
         according to the weight.
9]     If t appears in d and not member in ITS
         Then Add t to ITS.
10]    Remove t from TS'.
11]  Until t appears in d.
12]  Remove d from DS.
13]  Assign TS to TS'.
14] Until DS is empty.
15] Return ITS.
```

[그림 4] 초기 대표 용어 선택 알고리즘

4.1 소수의 긍정적 문서 집합에 대한 성능 평가

4.1.1 실험 환경

실험 문서 집합으로는 Reuters-21578을 선택하였다 [13]. 본 논문에서는 Reuters-21578의 TOPICS 범주들을 선택하였으며 ApteMod 버전을 사용했고 라벨이 없는 문서들은 제외시켰다. 실험 대상으로 소수 예제 문서 집합들을 준비하고자 테스트 문서 집합과 학습 문서 집합에 적어도 하나의 문서를 각각 포함하고 있는 범주(category)들을 선택(총 90개)한 후 이중에서 학습 문서 개수가 10개~30개인 범주 21개를 마지막으로 선별했다.

테스트 문서 집합은 3019개의 문서들을 포함하고 있다. 용어의 역문헌 빈도수(IDF)값을 구하기

위해 90개의 범주들에 속하는 7770개의 학습 문서 집합으로부터 문서 빈도수 정보를 이용하였다. 사용자는 자신의 관심 사항에 부합하는 긍정적 문서 집합(positive documents)만을 제공한다는 가정하에 알고리즘 수행시 부정적 문서(negative documents)들의 정보 이용은 모두 제외시켰다.

비교 대상 알고리즘에 사용된 벡터들의 가중치는 용어의 TF×IDF로 계산하였다[1]. 실험시 사용된 조정 상수(parameter)들의 설정값들은 [1, 10]에서 사용한 값을 이용하였다. 즉, Rocchio의 경우 $\alpha=0, \beta=1, \gamma=0$로 두었고 Widrow-Hoff의 경우 $\eta=0.25$, y_i=1로 두었다. 유사도 계산식은 cosine 식 [12]을 이용했으며 문서 분류 성능의 척도로서 11 points average precision [3, 8, 12]를 사용했다. 그리고 식 5의 p 값으로 10 을 사용하였다.

4.1.2 실험 결과

RKEF, 즉 퍼지추론과 연관 피드백 방법을 결합시킨 방법의 결과는 표 1과 같다. 표 1을 살펴보면 RKEF이 다른 비교 방법들보다 나은 성능을 보여주고 있음을 확인할 수 있다. 실험에 사용한 문서 분류가 소수의 문서 (10 ~ 30개)로 구성되었음을 주지하기 바란다. 여기서, RO는 Rocchio 방법을 WH는 Widrow-Hoff의 방법을 의미한다.

4.2 다수의 긍정적 문서 집합에 대한 성능 평가

위 실험을 통하여 본 방법은 적당수 (10 ~ 30개)의 관련문서가 주어진 경우 좋은 성능을 보임을 확인할 수 있었다. 본 실험에서는 사용자로부터 더 많은 수의 관심 문서들이 주어진 경우를 고려하여 위 방법의 성능을 확인하고자 하였다. 우선 다수의 문서 집합에 대한 위 방법의 성능을 살펴본 후 문서들을 여러 개의 소수 문서 집합들로 분할한 뒤 각각에 대해 위 방법을 적용시키는 방법에 대한 성능을 살펴보았다.

문서 집합을 여러 개의 문서 집합들로 나누는 방법들에는 hierarchical clustering과 K-means 들이 있다 [14, 15]. 그러나 이들 방법을 사용하기에는 어려운 점들이 있다. 우선 전자의 경우 클러스터링 레벨을 미리 결정해야만 하고 후자의 경우 또한 분할되어 생성될 문서 그룹 즉 클러스터 개수 (K)를 미리 결정해야만 하는 제약들이 있다. 따라서 본 실험에서는 앞에서 열거한 일반적인 분류 방법을 사용하지 않고 문서 분류 쪽에서 좋은 성능을 보이고 있는 GIS(Generalized Instance Set) 방법에 본 방법을 적용시켰다.

[표 1] 21개 범주들에 대한 분류 성능

Category	11 points average precision		
	RO	WH	RKEF
lumber	0.346	0.354	0.550
dmk	0.044	0.042	0.084
sunseed	0.376	0.375	0.451
lei	0.273	0.273	0.363
soy-meal	0.539	0.447	0.772
fuel	0.429	0.436	0.518
soy-oil	0.185	0.185	0.323
heat	0.483	0.480	0.626
lead	0.556	0.557	0.614
housing	0.373	0.373	0.352
strategic-met	0.127	0.137	0.120
hog	0.513	0.533	0.485
orange	0.933	0.933	0.975
tin	0.959	0.966	0.986
rapeseed	0.443	0.428	0.575
wpi	0.764	0.708	0.728
pet-chem	0.405	0.482	0.308
silver	0.377	0.508	0.770
zinc	0.880	0.799	0.921
retail	0.030	0.024	0.194
sorghum	0.489	0.342	0.591
Average	0.454	0.447	0.538

4.2.1 실험 환경

실험 시 사용된 조정 상수는 앞 실험과 동일하다. 그렇지만, 실험 대상 범주는 앞 실험에서 설명한 90개의 범주에서 다수의 학습 문서들을 가진 상위 20개의 범주를 선택하였다. GIS 분류기를 이용한 실험에서는 GIS 알고리즘의 일반화 함수(Generalization function)에 Rocchio, Widrow-Hoff 그리고 본 제안 방법을 적용시킬 수 있다. GIS 알고리즘의 일반화 함수에 사용된 k 값으로 10 에서 150 사이의 10 단위로 선택한 총 15개의 k 값들을 선택하고 각각에 대해 분류 실험을 수행하였다.

4.2.2 실험 결과

표 2는 다수의 학습 문서를 가진 상위 20 개의 범주들에 대한 문서 분류 성능을 보여주고 있다. 결과를 살펴보면 퍼지 추론과 연관 피드백 방법을 결합시킨 방법(RKEF)의 성능이 Rocchio 와 Widrow-Hoff 의 성능들과 큰 차이를 보여주고 있지 않다. 이는 다수의 문서 집합에 대해서는 문서들을 그룹핑한 후 각각의 그룹에 대해서 RKEF 방법을 적용해야 함을 의미한다. 이러한 목적으로 본 논문에서는 GIS 알고리즘과 RKEF 방법을 결합시켜 사용하였다.

[표 2] 다수의 학습 문서를 가진 상위 20 개의 범주들에 대한 성능

Category	11 points average precision		
	RO	WH	RKEF
nat-gas	0.492	0.494	0.591
soybean	0.639	0.589	0.708
veg-oil	0.626	0.630	0.625
gold	0.855	0.843	0.831
gnp	0.816	0.820	0.888
coffee	0.936	0.979	0.951
oilseed	0.483	0.425	0.434
sugar	0.739	0.776	0.72
dlr	0.636	0.686	0.698
money-suppl	0.334	0.587	0.697
corn	0.644	0.624	0.677
ship	0.822	0.745	0.781
wheat	0.764	0.798	0.802
interest	0.636	0.720	0.641
trade	0.717	0.661	0.705
crude	0.778	0.801	0.791
grain	0.802	0.871	0.837
money-fx	0.582	0.537	0.613
acq	0.576	0.727	0.718
earn	0.961	0.948	0.908
Average	0.692	0.713	0.731

표 3은 GIS 알고리즘의 일반화 함수에서 긍정적 문서들만을 사용했을 경우, 위의 k 값들에 대한 실험 결과들 중에서 일반화 과정에 사용된 방법별로 각 범주별 가장 좋은 성능값들을 선택해서 보여주고 있다. 표 3의 결과를 살펴보면 GIS 알고리즘의 일반화 함수에서 긍정적 문서들만을 사용했을 경우 다른 비교 방법들에 비해서 RKEF가 향상된 성능을 보여주고 있음을 확인할 수 있다.

표 4는 다수의 학습 문서 집합을 가진 상위 20 개의 범주들에 대하여 RKEF만을 사용했을 경우의 성능과 GIS 방법rhk 결합했을 경우의 성능을 함께 보여주고 있다. 결과를 살펴보면 다수의 문서 집합에 퍼지 추론 기반 방법을 소수의 문서 집합에 적용할 때와 마찬가지로 동일하게 그대로 적용하기보다는 다수의 문서 집합을 여러 개의 소수의 문서 집합들로 그룹화하는 GIS 방법과 함께 사용하는 것이 보다 효과적임을 알 수 있다.

[표 3] 긍정적 문서만을 일반화에 사용한 경우의 성능

Category	ROP	WHP	Best		
			GISRP	GISWP	GISFP
nat-gas	0.492	0.494	0.518	0.599	0.643
soybean	0.639	0.589	0.642	0.654	0.738
veg-oil	0.626	0.630	0.657	0.651	0.756
gold	0.855	0.843	0.861	0.863	0.846
gnp	0.816	0.820	0.831	0.835	0.871
coffee	0.936	0.979	0.947	0.936	0.989
oilseed	0.483	0.425	0.497	0.508	0.601
sugar	0.739	0.776	0.793	0.807	0.882
dlr	0.636	0.686	0.719	0.726	0.751
money-suppl	0.334	0.587	0.624	0.607	0.726
corn	0.644	0.624	0.658	0.654	0.797
ship	0.822	0.745	0.831	0.821	0.854
wheat	0.764	0.798	0.808	0.803	0.861
interest	0.636	0.720	0.731	0.738	0.793
trade	0.717	0.661	0.733	0.740	0.749
crude	0.778	0.801	0.809	0.808	0.846
grain	0.802	0.871	0.859	0.866	0.867
money-fx	0.582	0.537	0.616	0.615	0.663
acq	0.576	0.727	0.707	0.708	0.792
earn	0.961	0.948	0.963	0.962	0.962
Average	0.692	0.713	0.740	0.745	0.799

4.3 다수의 부정적 문서를 고려한 경우의 성능 평가

RKEF 방법은 처음부터 긍정적 문서만 주어진다는 가정 하에 제안되었다. 하지만, 본 논문에서는 퍼지 추론 방법과 연관 피드백 방법을 그대로 유지한 채 부정적 문서까지 고려할 경우의 성능 평

가를 시도하였다. 식 6은 학습 문서 집합 중에서 부정적 문서집합을 해당 범주의 대표 벡터를 구성하는데 사용하기 위해서 본 실험에서 사용된 계산식이다. 핵심적인 초기 대표 용어 집합을 퍼지 추론을 통해서 긍정적 문서 집합과 부정적 문서 집합으로부터 개별적으로 추출하여 용어 가중치 재산정에 사용하였다. GIS 알고리즘을 적용하여 클러스터를 생성할 경우 그 클러스터 안에는 일반적으로 긍정적 문서와 부정적 문서를 모두 포함하며 문서의 수는 몇몇 클러스터를 제외하고는 인수로 주어진 k가 된다.

$$w_j^C = \alpha \frac{\sum_{i \in P} w_{ij} \times RD_{ij}^{K_P}}{n_p} - \beta \frac{\sum_{i \in N} w_{ij} \times RD_{ij}^{K_N}}{n_n} \qquad \text{(식 6)}$$

여기서, P는 주어진 클러스터 C 내의 긍정적 문서의 집합을, N은 부정적 문서의 집합을, n_p와 n_n은 각각 그 집합의 문서의 개수를 의미한다. 그리고, K_P는 P에 퍼지추론 방법을 적용시켜 추출한 초기 대표 용어 집합을, K_N은 N에서 추출한 초기 대표 용어 집합을 의미한다. 또한, w_j^G 는 클러스터 C에서의 용어 j 의 가중치를, w_{ij}는 문서 i 에서 용어 j 의 가중치를, $RD_{ij}^{K_P}$ 는 문서 i 에서 용어 j 와 K_P 간의 관련정도를, $RD_{ij}^{K_N}$ 는 문서 i 에서 용어 j 와 K_N 간의 관련정도를 나타낸다.

4.3.1 실험 환경

실험 대상 범주는 4.2 실험과 동일하다. 앞의 모든 실험은 긍정적 문서만 고려한 실험이어서 기존 방법과 비교시 사용된 조정 상수(parameter)들의 설정값들은 Rocchio의 경우 α = 0, β = 1, γ = 0, Widrow-Hoff의 경우 η = 0.25 이었다. 따라서, 부정적 문서들을 포함한 실험에서는 이를 변경하여 사용하였다. 즉, Rocchio의 경우 α = 0, β = 1, γ = 1, Widrow-Hoff 의 경우 η = 0.25를 사용하였다. 그리고, 본 제안 방법(식 6) 의 경우 α = 1, β = 1 로 두어 실험하였다.

4.3.2 실험 결과

표 5는 GIS 알고리즘의 일반화 함수에서 부정적 및 긍정적 문서들을 사용했을 경우, k 값들에 대한 실험 결과들 중에서 일반화 과정에 사용된 방법별로 각 범주별 가장 좋은 성능값들을 선택해서 보여주고 있다.

표 5의 결과를 살펴보면 GIS-W(GIS + Widrow-Hoff)가 GIS-R (GIS+Rocchio) 보다 성능이 낮고 GIS-R와 GIS-W의 성능과 RO (Rocchio)와 WH (Widrow-Hoff)의 성능사이에 큰 차이를 보여주지 않고 있다. 이것은 방법별로 최적의 k 값을 찾아서 사용하지 않았기 때문이다. 그러나 이러한 다소 제약된 실험환경에서도 퍼지 추론을 이용한 방법의 유용성을 충분히 확인할 수 있을 것으로 판단되어 방법별 최적의 k 값들을 찾는 실험은 제외하였다.

[표 4] RKEF 만을 적용했을 경우 성능과 GIS 방법에 적용했을 경우의 성능

Category	11 points average precision	
	RKEF	GIS-FP
nat-gas	0.591	0.643
soybean	0.708	0.738
veg-oil	0.625	0.756
gold	0.831	0.846
gnp	0.888	0.871
coffee	0.951	0.989
oilseed	0.434	0.601
sugar	0.72	0.882
dlr	0.698	0.751
money-suppl	0.697	0.726
corn	0.677	0.797
ship	0.781	0.854
wheat	0.802	0.861
interest	0.641	0.793
trade	0.705	0.749
crude	0.791	0.846
grain	0.837	0.867
money-fx	0.613	0.663
acq	0.718	0.792
earn	0.908	0.962
Average	0.731	0.799

GIS 알고리즘의 일반화 함수에서 부정적 문서들

을 제외한 긍정적 문서들만을 사용했을 경우는 표 3에서 보는 바와 같이 RKEF가 다른 비교 방법들에 비해서 향상된 성능을 보여주고 있다. 그러나 RKEF 방법을 GIS에 적용한 결과 (GIS-F)는 Rocchio를 GIS 에 적용한 결과(GIS-R)에 비해서 별다른 성능 향상을 보여주지 않고 있다. 이는 퍼지 추론에 사용된 규칙과 멤버쉽 함수가 긍정적 문서만을 고려해서 만들어진 것인데, 이를 그대로 부정적 문서에도 적용했기 때문으로 보인다.

5. 결 론

본 논문에서는 소수의 긍정적 문서 집합을 대상으로 문서들의 내용을 대표하는 중요 용어들을 추출하고 이들의 가중치를 부여하는 문제를 해결하기 위한 방법인 퍼지 추론 및 용어 발생 빈도수의 유사성을 이용한 가중치 재산정 접근 방법을 GIS 알고리즘에 적용시켜 문서분류 성능을 비교해 보았다. GIS 알고리즘에 적용시켜 봄으로써 소수 학습 문서 집합을 대상으로 한다는 제약성을 극복할 수 있었으며 긍정적 문서들만을 일반화에 사용한 실험에서 나은 성능을 보여줌으로써 성능 향상의 가능성을 확인 할 수 있었다. 향후 부정적 문서 집합을 고려한 퍼지 추론 방법에 대한 연구가 진행된다면 더 나은 성능을 기대할 수 있을 것으로 보인다.

[표 5] 긍정적 및 부정적 문서를 일반화에 함께 사용한 경우의 성능

Category	RO	WH	Best		
			GIS-R	GIS-W	GIS-F
nat-gas	0.567	0.633	0.723	0.694	0.667
soybean	0.576	0.758	0.78	0.74	0.779
veg-oil	0.568	0.703	0.739	0.716	0.701
gold	0.833	0.8	0.862	0.866	0.853
gnp	0.743	0.914	0.932	0.915	0.93
coffee	0.929	0.964	0.988	0.991	0.985
oilseed	0.553	0.723	0.663	0.665	0.613
sugar	0.694	0.889	0.91	0.914	0.917
dlr	0.682	0.705	0.805	0.777	0.787
money-suppl	0.412	0.676	0.726	0.725	0.727
corn	0.661	0.821	0.898	0.857	0.901
ship	0.865	0.876	0.88	0.866	0.876
wheat	0.732	0.845	0.893	0.874	0.929
interest	0.695	0.718	0.803	0.79	0.802
trade	0.735	0.769	0.788	0.77	0.81
crude	0.804	0.857	0.88	0.838	0.892
grain	0.779	0.926	0.937	0.96	0.946
money-fx	0.581	0.749	0.694	0.686	0.688
acq	0.885	0.911	0.877	0.822	0.875
earn	0.953	0.954	0.967	0.966	0.964
Average	0.712	0.810	0.837	0.819	0.832

참 고 문 헌

[1] D.D.Lewis, R.E.Schapore, J.P.Call, and R.Papka. "Training algorithms for linear text classifiers", *Proc. of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 298-306, 1996.

[2] 노순억, 김병만, 허남철, "퍼지 추론을 이용한 소수 문서의 대표 키워드 추출", *한국퍼지 및 지능시스템학회 논문지*, vol. 11, No. 9, pp. 837-843, 2001.

[3] Sebastiani, F. "Machine Learning in Automated Text Categoeisation", *Technical Report IEI-B4-31-1999, Istituto di Elaborazione dellInformazione*, 1999.

[4] Tom M. Mitchell. *Machine Learning*, McGraw-Hill, 1997.

[5] Y. Yang, "Expert network: effective and efficient learning from human decisions in text categorization and retrieval", *Proc. of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, 1994.

[6] T. Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proc. of European Conference on Machine Learning*, pp. 137-142, 1998.

[7] Y. Yang and X. Liu. "A re-examination of text categorization methods", *Proc. of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999.

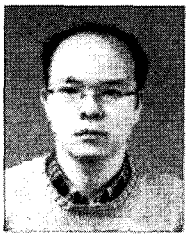
[8] Y. Yang. "An evaluation of statistical approaches to text categorization". *Journal of Information Retrieval*, pp. 67-88, 1999.
[9] L. S. Larkey and W. B. Croft. "Combining classifiers in text categorization", *Proce. of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 289-297, 1996.
[10] K. Lam and C. Ho, "Using a generalized instance set for automatic text categorization", *Proc. of 21th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 88-89, 1998.
[11] Kwok-Yin Lai and Wai Lam, "Automatic Textual Document Categorization Using Multiple Similarity-Based Models", *Proc. of SDM01*, 2001.
[12] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, NY, USA, 1999.
[13] http://www.research.att.com/~lewis/reuters21578.html
[14] William B. Frakes, Ricardo Baeza-Yates, *Information Retrieval : Data Structures & Algorithms*, Prentice Hall, pp.102-160, 1992.
[15] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques", *Proc. of KDD Workshop on Text Mining*, 2000.

노 순 억 (Sun-Ok Rho)

- 1999년 : 금오공과대학교 컴퓨터공학과 학사
- 2003년 : 금오공과대학교 컴퓨터공학과 공학석사
- 관심분야 : 정보검색, 인공지능
- E-mail: sorho@se.kumoh.ac.kr

김 병 만 (Byeong Man Kim)

- 1987년 : 서울대학교 컴퓨터공학과 학사
- 1989년 : 한국과학기술원 전산학과 공학석사
- 1992년 : 한국과학기술원 전산학과 컴퓨터공학 박사
- 1992년~ 현재 : 금오공과대학교 컴퓨터 교수
- 1998년~1999년 : 미국 Univ. of California, Irvine Post Doc.
- 관심분야 : 인공지능, 정보검색, 소프트웨어 검증
- E-mail : bmkim@se.kumoh.ac.kr



오 상 엽 (Sang Yeop Oh)

- 1992년 : 한국과학기술원 물리학 학사
- 1994년 : 한국과학기술원 전산학 석사
- 2001년 : 한국과학기술원 전자전산학 전산학 전공 박사
- 2001년~2002년 : (주) 서치솔루션 선임연구원
- 2002년~2003년 : University of Michigan 방문연구원
- 2003년~2004년 : 한국과학기술원 초빙교수
- 2004년~현재 : 금오공과대학교 컴퓨터공학부 전임강사
- 관심분야 : 정보검색, 인공지능
- E-mail: syoh@kumoh.ac.kr

이 현 아 (Hyun Ah Lee)

- 1996년 : 연세대학교 컴퓨터과학과 학사
- 1998년 : 한국과학기술원 전산학과 석사
- 2004년 : 한국과학기술원 전산학과 박사
- 2000년~2004년 : ㈜ 다음소프트 자연언어처리연구소 팀장
- 2004년~현재 : 금오공과대학교 컴퓨터공학부 전임강사
- 관심분야 : 자연언어처리 정보검색 지식공학 기계번역
- E-mail : halee@kumoh.ac.kr